2003

Entropy and generalized least square methods in assessment of the regional value of streamgages

Momcilo Markus
Illinois State Water Survey, momcilo@sws.uiuc.edu

H. Vernon Knapp
Illinois State Water Survey, vknapp@sws.uiuc.edu

Gary D. Tasker
U.S. Geological Survey, gdtasker@usgs.gov

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Hydrology 283 (2003) 107–121
Journal of Hydrology
www.elsevier.com/locate/jhydrol

Entropy and generalized least square methods in assessment of the regional value of streamgages

Momcilo Markus[a,*], H. Vernon Knapp[a,1], Gary D. Tasker[b,2]

[a]*Illinois State Water Survey, Watershed Science Section, 2204 Griffith Drive 518, 61820 Champaign, IL, USA*
[b]*US Geological Survey, 430 National Center, 20192 Reston, VA, USA*

Received 11 December 2001; accepted 24 June 2003

Abstract

The Illinois State Water Survey performed a study to assess the streamgaging network in the State of Illinois. One of the important aspects of the study was to assess the regional value of each station through an assessment of the information transfer among gaging records for low, average, and high flow conditions. This analysis was performed for the main hydrologic regions in the State, and the stations were initially evaluated using a new approach based on entropy analysis. To determine the regional value of each station within a region, several information parameters, including total net information, were defined based on entropy. Stations were ranked based on the total net information. For comparison, the regional value of the same stations was assessed using the generalized least square regression (GLS) method, developed by the US Geological Survey. Finally, a hybrid combination of GLS and entropy was created by including a function of the negative net information as a penalty function in the GLS. The weights of the combined model were determined to maximize the average correlation with the results of GLS and entropy. The entropy and GLS methods were evaluated using the high-flow data from southern Illinois stations. The combined method was compared with the entropy and GLS approaches using the high-flow data from eastern Illinois stations.

Keywords: Illinois; Gaging network; Entropy; Generalized least square method; Information transfer; Network design

1. Introduction

The Illinois streamgaging network, operated by the US Geological Survey (USGS) provides hydrologic information that is vital for the management of the State's water resources. Streamflow data are used in forecasting floods; for evaluation of watershed and stream management practices; and for planning and operation of dams and reservoirs, water supplies, and wastewater treatment facilities. In addition, streamflow data are used to develop regional relationships for the analysis of the frequencies and probabilities of low and high flows, and to provide input to numerical models. The total network in 2002 included 186 stations: 154 continuous discharge stations, 23 continuous stage-only stations and 9 crest-stage stations.

* Corresponding author. Tel.: +1-217-333-0237; fax: +1-217-333-2304.
E-mail addresses: momcilo@sws.uiuc.edu (M. Markus), v-knapp@sws.uiuc.edu (H. Vernon Knapp), gdtasker@usgs.gov (G.D. Tasker).
[1] Tel.: +1-217-333-4423; fax: +1-217-333-2304.
[2] Tel.: +1-703-648-5892; fax: +1-703-648-5484.

0022-1694/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0022-1694(03)00244-0

Nomenclature

β	$-(p \times 1)$ vector of parameters in generalized least square method
Λ	$-(N \times N)$ error covariance matrix in generalized least square method
$C(i)$	−the station cost for station i
e	$-(N \times 1)$ vector of random errors in generalized least square method
$H(X)$	−entropy of variable X
$H(X, Y)$	−joint entropy of variables X and Y
$H(Y\|X)$	−conditional entropy of Y given X
H, H_r	−leverage-related statistics to rank the stations in generalized least square method
k	−discrete data interval for variable X in entropy computations
K	−number of class intervals (possible outcomes) for X in discrete entropy computations
L	−number of class intervals for Y in discrete entropy computations
l	−discrete data interval for variable Y in entropy computations
M	−variable multiplier for the cost function in the combined model
$N(i)$	−total net information at station i
N	−number of sites in generalized least square method
$p(x_k)$	−the probability of x_k in discrete entropy computations
$p(x_k, y_l)$	−probability of an outcome corresponding to interval k for X and interval l for Y
p	−number of parameters in generalized least square method
$R(X, Y)$	−information received by station X from station Y
$R(i)$	−information received at station i, from all other stations
$S(X, Y)$	−information sent from station X to station Y
$S(i)$	−information sent by station i to all other stations
$T(X, Y)$	−transinformation between X and Y
x_k	−an outcome corresponding to interval k in discrete entropy
X_r	−matrix of the basin characteristics for a representative set of stations in a region
X	$-(N \times p)$ matrix of $(p - 1)$ basin characteristics augmented by a column of ones
$\hat{Y}$	$-(N \times 1)$ vector of estimated flow characteristics at N sites
Y	$-(N \times 1)$ vector of observed flow characteristics in generalized least square method

First, two different approaches, entropy analysis and the generalized least square (GLS) method were used to assess the value of each station in regional analyses. The entropy approach evaluates stations through their information transmission to and from other stations, as defined in information theory. A network analysis based on the GLS method was developed by the USGS and incorporated into the existing computer program for regional regression analysis (GLSNET). Once a regional regression equation is derived, and the planning horizon defined, stations are ranked based on several factors including station cost, record length, cross correlation between stations, and station leverage representing a function of various physical parameters such as stream slope and drainage area. Next, a combined method, which uses both entropy and GLS methods, was created to combine the assumptions of the two approaches into a single method. The combined method relies on GLSNET and uses a simple linear function of net entropy as a penalty function. A purpose of this study is to evaluate the performance of the combined method.

The entropy approach used in this study seeks to determine the best stations to discontinue and also the best stations to continue in a gage network. The GLS approach is a station-continuous approach that also seeks to determine the best stations (either existing or planned) to continue, subject to specified constraints, into a planning horizon. For method comparison in this study, starting new stations was not an option, and the entropy, GLS, and combined methods were compared and contrasted using only existing stations.

2. Network design

Two basic approaches have been used in previous network evaluation studies: (1) statistical analyses to determine the value of the network and individual gages for regional regression, and (2) qualitative evaluations of the use, characteristics, and overall value of the network. Although both types of approaches are necessary for a comprehensive network evaluation, this paper will focus on the first approach.

Streamgaging network design has been an active area of research since the 1960s. An optimal interpolation developed by Gandin (1965) was used to specify a minimum spatial density of observation stations for a given accuracy of estimation. Karasev (1968) proposed a technique, which specifies a range in the number of stream-gages required to estimate runoff within a given accuracy. Moss and Karlinger (1974) outlined the Network Analysis for Regional Information (NARI) strategy for network design based on Monte–Carlo simulation, regression analysis and Bayesian analysis. Moss et al. (1982) described the NARI technique in more detail. Mades and Oberg (1986) applied the NARI procedure to evaluate the streamgaging network in Illinois and also applied Kalman-filtering analysis to evaluate cost-effective stream gaging strategies using a method outlined in Moss and Gilroy (1980). Network analysis using GLS as described by Tasker (1986) was used to identify an efficient gaging plan for a specific operating budget and provide insight about the extent of regional information lost or gained by reducing or increasing the operating budget. Moss and Tasker (1991) compared GLS and NARI techniques and demonstrated the better performance of GLS-based network analysis over a wide range of data availabilities and design constraints.

National Research Council (1992) assessed the regional hydrology and the USGS stream-gaging network. Among several other conclusions and recommendations, the Council stated that the assumption of data stationarity is not appropriate. Thomas (1994) presented a review of the techniques used in data networks and provided future recommendations. Efforts need to be expended in the future in developing water-quality network analysis techniques and techniques for the coordinated analysis of meteorological, water quality and stream flow networks. Wahl and Thomas (1995) described current and future needs of the USGS stream-gaging program, and provided particularly detailed specific categories of use. USGS (1999) presented the thoughts of the USGS regarding the future of stream flow information for the Nation. The paper seeks input from all interested parties regarding the USGS' vision of National Stream flow Information Program.

3. Entropy

The concept of entropy has been very popular in the scientific literature over the last several decades. Entropy, as defined in information theory, is a measure of uncertainty of a particular outcome in a random process, and provides an objective criterion in selecting the mathematical model. Linfoot (1957) demonstrated that the advantage of using informational correlations in physical applications is that they are invariant under transformations, which is not the case with an ordinary correlation. Amorocho and Espildora (1973) and Valdes et al. (1975) were among the first to introduce the basics of entropy in hydrology. Harmancioglu et al. (1986) compared correlation-based measures and the entropy-based measures of information transfer between variables. Several ways to improve the information transfer between two sets of variables were addressed. They also discussed additional advantages and disadvantages of the entropy-based approach, pointing out that the entropy principle does not assume normality or any particular type of functional relationship (linear or non-linear).

Entropy-based techniques also have been used in various studies for gage network design. Husain (1989) expressed the information-transmitting capabilities of a hydrologic network in terms of entropy and proposed a gage network design method based on entropy. Harmancioglu and Alpaslan (1992) used the information-based uncertainty measure in water-quality monitoring network design. Yang and Burn (1994) described an entropy-based approach to design streamgaging network. Their method is based on a directional informational transfer (DIT) index, which their study favourably compared with the traditional correlation coefficient. Yang and Burn (p. 308) state

that "Entropy and mutual information possess advantages relative to other measures of association in that they provide a quantitative measure of: (1) the information at a station; (2) the information transferred and lost during the transmission; (3) a description of the relationships among stations according to their information transmission characteristics."

A discrete form of entropy is given by (Press et al., 1995):

$$H(X) = \sum_{k=1}^{K} p(x_k) \log \frac{1}{p(x_k)} \tag{1}$$

where k denotes a discrete data interval, x_k is an outcome corresponding to interval k, and $p(x_k)$ is the probability of x_k. The probability $p(x_k)$ is based on the empirical frequency of variable X. The entropy is expressed in bits because the base of the logarithm was assumed to be equal to 2. Variable X can have only K outcomes. For continuous variables, such as stream flow discharge, a finite number of class intervals K must be chosen. Entropy $H(X)$ is also called marginal entropy of a single variable X. Uncertainty of two variables, X and Y, can be described by joint entropy $H(X, Y)$. Joint entropy is defined by:

$$H(X, Y) = \sum_{k=1}^{K} \sum_{l=1}^{L} p(x_k, y_l) \log \frac{1}{p(x_k, y_l)} \tag{2}$$

where k denotes a discrete data interval for variable X, l denotes a discrete data interval for variable Y, $p(x_k, y_l)$ is a probability of an outcome corresponding to interval k for X and interval l for Y, K is the number of class intervals (possible outcomes) for X, and L is the number of class intervals for Y. Our applications assumed $K = L$. Classes were based on range of values for discharge on a site.

The joint and marginal entropies are related:

$$H(X, Y) = H(X) + H(Y) - T(X, Y) \tag{3}$$

where $T(X, Y)$ is the information transferred from X to Y, called transinformation. Transinformation is a reduction of the original uncertainty, and it can be viewed as information about a predicted variable transferred by the knowledge of a predictor. Transinformation $T(X, Y)$ can be computed as follows:

$$T(X, Y) = H(X) + H(Y) - H(X, Y) \tag{4}$$

It can be shown that the transinformation is symmetrical, $T(X, Y) = T(Y, X)$ If X and Y are independent, $T(X, Y) = 0$, and $H(X, Y) = H(X) + H(Y)$. Uncertainty of Y, given X, denoted as $H(Y|X)$, is equal to:

$$H(Y|X) = H(Y) - T(X, Y) \tag{5}$$

If X and Y are independent, $H(Y|X) = H(Y)$, which means that the entropy of Y after obtaining X is the same as the original (marginal) entropy of Y. On the other hand, if the knowledge of X gives complete information about Y, the conditional entropy $H(Y|X)$ is equal to zero, and $H(Y) = T(X, Y)$. Transinformation can also be expressed as:

$$T(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{6}$$

Similar to the idea of directional information transfer (Yang and Burn, 1994), we define a fractional reduction of entropy of X by $R(X, Y)$:

$$R(X, Y) = \frac{T(X, Y)}{H(X)} \tag{7}$$

which also can be viewed as a reduction of uncertainty of X if Y is known, or information *received* by X from Y. Similarly, the information *sent* (transmitted) from X to Y is defined as:

$$S(X, Y) = \frac{T(X, Y)}{H(Y)} \tag{8}$$

The above equations describe the relationships between two variables, X and Y. The same reasoning can be applied to the network of streamgages. Using Eqs. (7) and (8), we define information received and sent at station i as:

$$R(i) = R(X(i), \hat{X}(i)) \tag{9}$$

where $X(i)$ represents the data at site i and the quantity $\hat{X}(i)$ at station i was obtained by linear regression:

$$\hat{X}(i) = a(i) + Y(i) \times b(i) \tag{10}$$

where $Y(i)$ is a matrix of data from all other stations, $a(i)$, $b(i)$ parameters of the regression between site i and all other sites. As the relations between the discretized (bin) data at different sites were found to be linear or close to linear, this assumption of linearity was deemed appropriate.

Similarly:

$$S(i) = S(X(i), \hat{X}(i)) \tag{11}$$

In this study the concept of entropy was used to determine the stations with the highest amounts of $S(i)$

and $R(i)$. A station denoted as i receives a lot of information if $R(i)$ is large relative to other stations. On the other hand, stations sending more information, having larger $S(i)$, are considered more valuable and will remain active. Finally, we define the net information transfer, $N(i)$, as the difference between $S(i)$ and $R(i)$:

$$N(i) = S(i) - R(i) \tag{12}$$

The value of each station can be quantified using $N(i)$. Stations with positive $N(i)$ are considered more valuable in regional analyses. If the number of stations in the network is to be reduced, such a station is more likely to be retained in the network than a station with a negative $N(i)$.

4. Generalized least square method

The GLS method was developed by the USGS to estimate regional regression equations to predict flow characteristics at ungaged sites. This analysis assigns different weights to the observed flow characteristics based on the record length, cross correlation with other sites, and the model error structure. The method is documented in Stedinger and Tasker (1985, 1986) and Tasker and Stedinger (1989).

The GLS method assumes that the suitable statistics, such as T-year peak flow, can be written as:

$$\hat{Y} = X\beta + e \tag{13}$$

where $\hat{Y}$ is a $(N \times 1)$ vector of estimated flow characteristics at N sites, X is an $(N \times p)$ matrix of $(p - 1)$ basin characteristics augmented by a column of ones, β represents a $(p \times 1)$ vector of parameters, and e is an $(N \times 1)$ vector of random errors. The dependent variable $\hat{Y}$ can be the logarithm of a 100-year flood, derived from a sample of logarithms of observed annual peak discharges at each site. The GLS estimator of β was suggested by Stedinger and Tasker (1985):

$$\beta = (X^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}X)^{-1}X^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}Y \tag{14}$$

where Y represents the $(N \times 1)$ vector of observed flow characteristics, and Λ represents the $(N \times N)$ error covariance matrix.

Sites from which to collect future streamflow data are identified with a mathematical program using regional information and are subject to budget constraints. An approximate solution is obtained using a step-backward technique that identifies those gaging stations that should be operated to maximize regional information as measured by the standard errors of prediction averaged over a representative set of ungaged sites in the region. A diagnostic statistic, leverage has a key role in ranking the stations. This statistic identifies points that are potentially influential due to their location in the regressor variable space. For station i, leverage is measured by the ith diagonal element of H, where:

$$H = X(X^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}X)^{-1}X^{\mathrm{T}}\boldsymbol{\Lambda}^{-1} \tag{15}$$

The GLS method actually uses a leverage-related statistic to rank the stations:

$$H_r = X_r(X^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}X)^{-1}X_r^{\mathrm{T}} \tag{16}$$

where X_r is a matrix of the same basin characteristics as X, but for a representative set of stations in a region. For this comparison study, X and X_r are identical. Thus, we assume that the basin characteristics for the gaged sites are representative of all sites in the region. For two stations with equal record length and equal operating costs, the station with higher leverage is considered more influential and thus more cost effective to operate in the future. Those stations having more unique physical characteristics will have high leverage and will often be considered to have more regional value. For any assumed planning horizon, the program will compute the rank of each station according to its importance, cost, and various physical parameters such as location, stream slope, and drainage area. Remote stations will also have higher leverage and more likely will be retained in the network.

5. Application

For the purpose of regional analysis, the stream-gages were classified by hydrologic regions. Each hydrologic region is based on the physiographic divisions in Illinois, which are defined by Leighton et al. (1948) on the basis of past glacial activities and their impact on landform and stream development.

Results are presented herein for hydrologic regions East and South as shown in Fig. 1.

5.1. Number of discrete intervals in entropy analysis

Choosing a number of discrete intervals is a practical problem identified by Valdes et al. (1975) and Chapman (1986): the entropy itself changes when the number of intervals is changed. For various numbers of class intervals, K, the entropy parameters, $S(i)$, $R(i)$, and $N(i)$, were computed at 12 gaging stations in southern Illinois having continuous records between 1952 and 1998, and stations were ranked according to their performance. The results (Table 1) show the extent in which the discretization interval influences the station ranks.

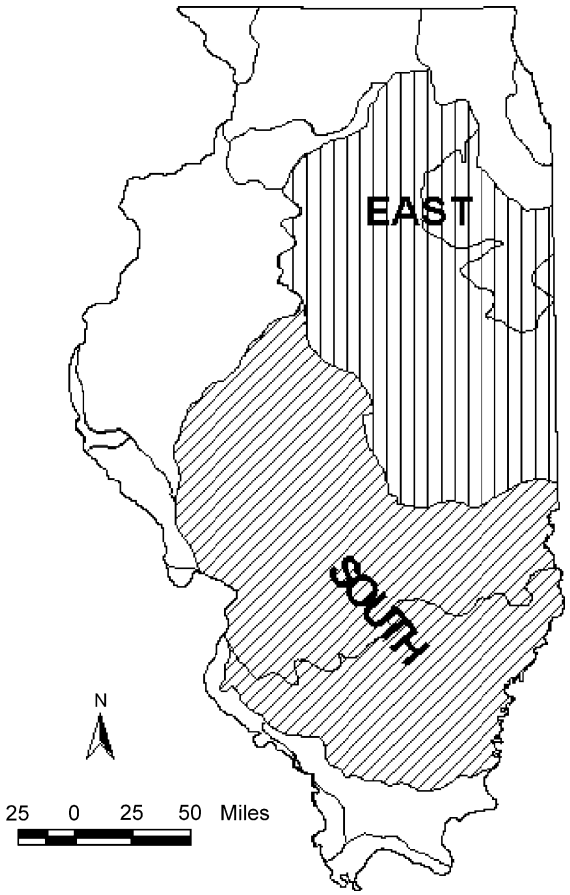

Fig. 1. Hydrologic regions East and South in Illinois.

All computation results are presented in Table 1, Fig. 2 (information sent), Fig. 3 (information received) and Fig. 4 (total net information). Although the entropy is sensitive to the changes in K, the ranks based on entropy appear to be less sensitive to K. For example, station 5 (USGS Gage #03380500) in Fig. 4 is consistently ranked as the first station to be discontinued for all values of K. Stations 2 (03346000) and 3 (03378000) are evaluated as stations with largest $N(i)$ so they would be least likely to be discontinued.

5.2. Comparing the performance of entropy analysis and GLS model

The performance of the entropy and GLS methods was tested on two examples using annual peak-flow data. The first set chosen for the analysis consists of 12 gaging stations in southern Illinois having continuous records between 1952 and 1998. The second set consists of the same gaging stations, but having only 20 years of continuous records (1979–1998).

The method based on entropy was applied using Eq. (1) for marginal entropy, Eq. (2) for joint entropy, and Eq. (4) for transinformation. Eqs. (7) and (9) were used to compute the total information received by a station i, $R(i)$. Eqs. (8) and (11) were used to calculate the total information sent by a station i, $S(i)$. Finally, Eq. (12) was used to compute the total net information transfer for station i, $N(i)$. Ten discrete class intervals were used in the above equations.

The GLS method was used to analyze the same stations for the same time period. However, instead of directly using the annual peak-flow data, the GLS method is based on a specific regional regression analysis using the annual data. In this example, the regression analysis for estimating the 25-year peak flood event was used. Watershed characteristics, considered independent variables in the regression equation, include the drainage area and channel slope. In the GLS analysis, the station rank was determined as a function of several independent variables including physical characteristics of the watershed, geographic location, planning horizon, and station cost.

The first example included the period 1952–1998. The information transfer parameters $S(i)$, $R(i)$, $N(i)$ as

Table 1
Rankings based on the information transmitted, $S(i)$, information received, $R(i)$ and net information, $N(i)$, for various numbers of class intervals, K, using annual peak flows for the period 1952–1998 in southern Illinois. Smaller numbers for ranks indicate that stations have less regional value

Station no.	USGS station no.	Number of class intervals											
		$K = 5$			$K = 10$			$K = 15$			$K = 25$		
		$S(i)$	$R(i)$	$N(i)$	$S(i)$	$R(i)$	$N(i)$	$S(i)$	$R(i)$	$N(i)$	$S(i)$	$R(i)$	$N(i)$
1	03345500	7	6	8	9	8	9	8	7	7	10	7	8
2	03346000	10	9	10	11	11	12	12	9	11	12	10	12
3	03378000	5	3	12	8	6	10	4	3	12	4	1	10
4	03379500	12	12	11	12	12	11	11	10	5	11	12	7
5	03380500	6	10	1	5	10	1	3	12	1	3	6	1
6	03381500	9	8	5	10	9	8	10	8	9	5	3	9
7	05576000	4	5	2	3	3	5	5	4	6	6	9	4
8	05577500	1	1	7	1	1	2	1	1	4	1	2	2
9	05587000	3	4	4	4	4	6	6	5	10	8	11	5
10	05588000	8	7	9	6	5	7	7	6	8	7	5	11
11	05594000	11	11	6	7	7	4	9	11	3	9	8	6
12	05597500	2	2	3	2	2	3	2	2	2	2	4	3

well as the station ranks based on those parameters are presented in Table 2, and Figs. 5 and 6. Lower ranks (for example, $r = 1$ or $r = 2$) indicate stations that are less important in the information transfer process and potentially could be the first stations to be discontinued should it become necessary to downsize the gaging network. Higher ranks ($r = 11$ or $r = 12$) indicate gages with high regional value that should be retained in the network. If a station ranked 1 were removed from the network, the station ranked 2 would not necessarily become the next rank 1 station. Removal of a gaging station can result in a different transinformation matrix in entropy, or leverage in GLS, producing a new, different set of station ranks. The correlation matrix for the rankings of information transmitted, $S(i)$, information received, $R(i)$, net

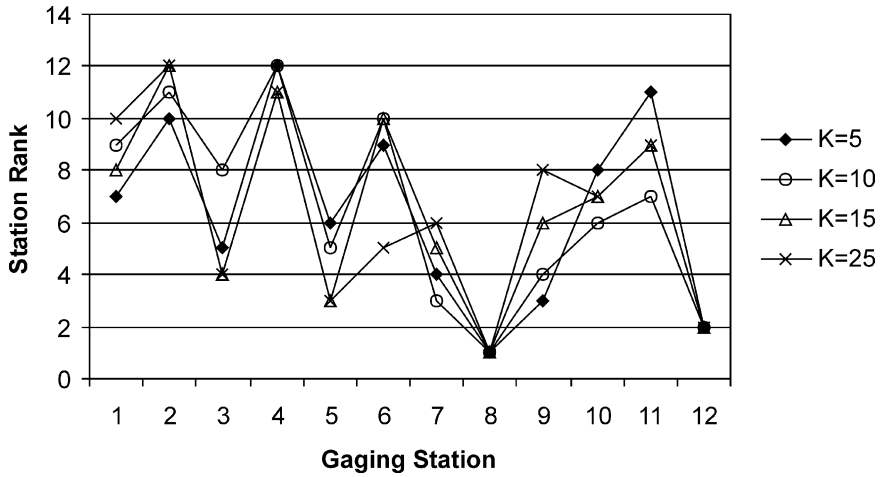

Fig. 2. Station rankings based on the total information transmitted at stations as a function of the number of discretization intervals using annual peak discharge time series for the period 1952–1998 in Southern Illinois.

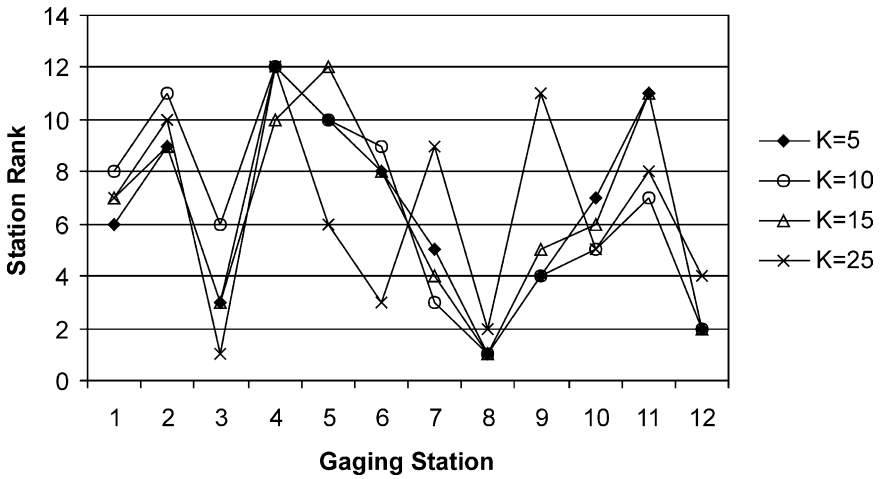

Fig. 3. Station rankings based on the total information received at stations as a function of the number of discretization intervals using annual peak discharge time series for the period 1952–1998 in Southern Illinois.

information, $N(i)$, and the GLS method, GLS(i), (Table 3) shows that information transmitted from one station to other stations is proportional to the information received at the same station from other stations. The net information correlates positively with both information received and information transmitted. The ranks based on GLS network analysis, denoted as GLS(i) and shown in Table 3, correlate negatively with the information received, $R(i)$, the information sent, $S(i)$ and total net information, $N(i)$.

The second example, uses 20 annual peak flows at the same stations, for the period 1979–1998. The GLS method again used the regional regression estimate for the 25-year flood, but this time it was based on only 20 years of data. The resulting parameters and their ranks are presented in Table 4, and Figs. 7 and 8. The correlation matrix between the ranks for each

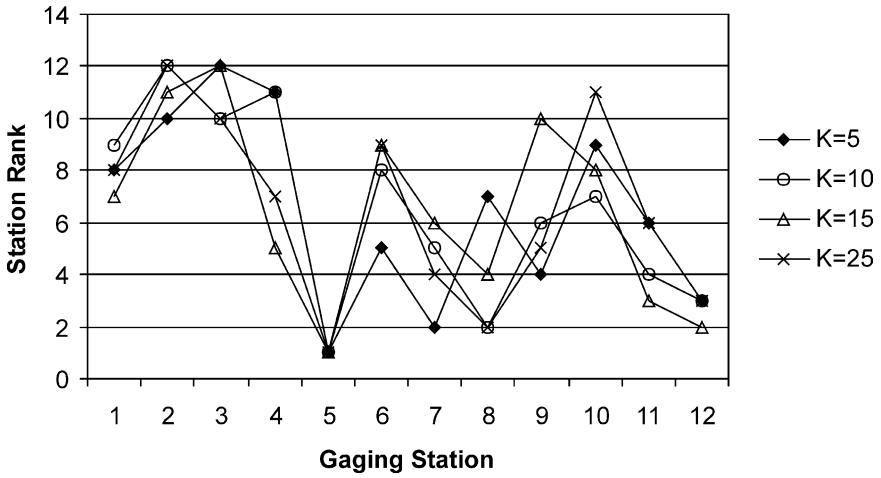

Fig. 4. Station rankings based on the total net information at stations as a function of the number of discretization intervals using annual peak discharge time series for the period of 1952–1998 in Southern Illinois.

Table 2
Information transmitted, $S(i)$, information received, $R(i)$, and net information, $N(i)$, and corresponding station rankings, including the generalized least square method, GLS(i), using annual peak flows for the period 1952–1998 in southern Illinois

Station no. (i)	USGS station number	Information			Rank (r)			
		$S(i)$	$R(i)$	$N(i)$	$S(i)$	$R(i)$	$N(i)$	GLS(i)
1	03345500	0.6002	0.5896	0.0106	9	8	9	2
2	03346000	0.6762	0.6412	0.0350	11	11	12	6
3	03378000	0.5777	0.5614	0.0163	8	6	10	9
4	03379500	0.6890	0.6680	0.0210	12	12	11	3
5	03380500	0.5329	0.6395	−0.1065	5	10	1	4
6	03381500	0.6004	0.5928	0.0076	10	9	8	8
7	05576000	0.4546	0.4677	−0.0131	3	3	5	7
8	05577500	0.3161	0.3377	−0.0216	1	1	2	10
9	05587000	0.4650	0.4761	−0.0112	4	4	6	1
10	05588000	0.5405	0.5476	−0.0071	6	5	7	11
11	05594000	0.5657	0.5847	−0.0190	7	7	4	5
12	05597500	0.3757	0.3959	−0.0203	2	2	3	12

method is shown in Table 5. The results show that the information transmitted from one station to other stations, correlates positively with the information received and with $N(i)$. The information received has a small negative correlation with $N(i)$. In this example, the information sent appears to be more dominant than the information received in $N(i)$. Finally, GLS(i) has a small negative correlation with $S(i)$, $R(i)$, and $N(i)$.

The results of this sensitivity test showing the station ranks as a function of the time period are presented in Table 6. The methods based on transmitted information, received information, and GLS appear to be less sensitive to the time period compared to the results based on net information.

5.3. The combined model

A weighted average between GLS and total net entropy $N(i)$ was created as an attempt to overcome the differences and combine the two approaches. The combined model is based on the USGS' computer

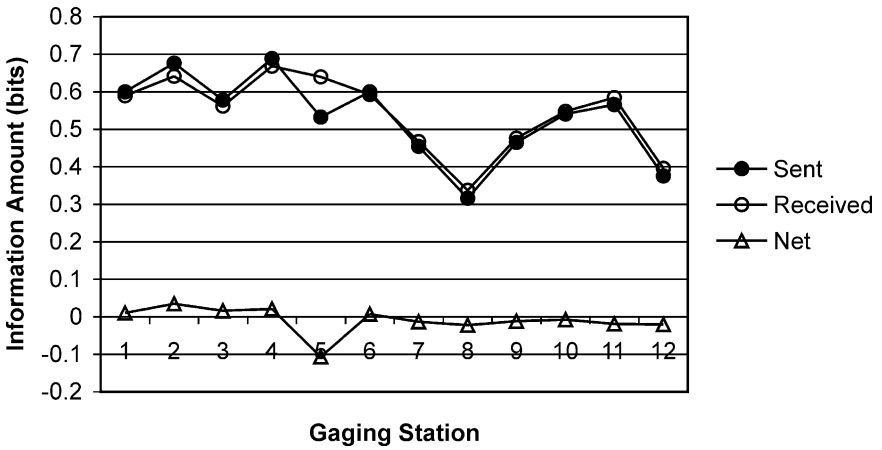

Fig. 5. Information transmitted, $S(i)$, information received, $R(i)$, and net information, $N(i)$, using annual peak discharge time series for the period 1952–1998 in southern Illinois. Gaging station numbers correspond to those in Table 2.

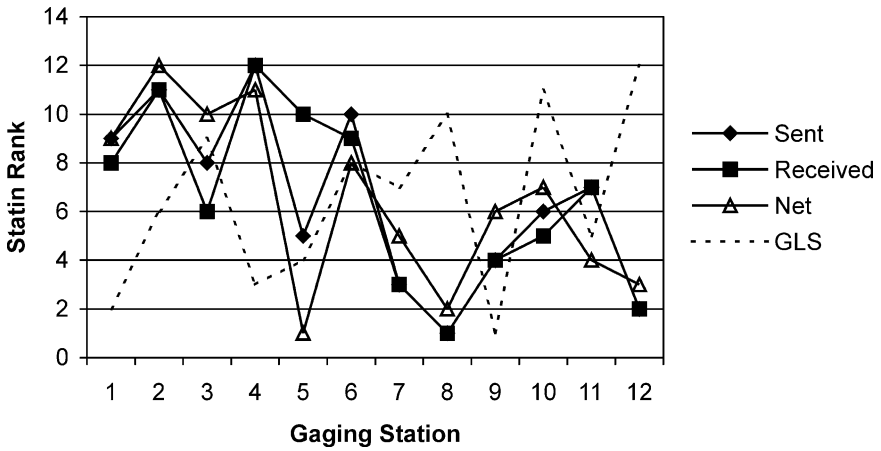

Fig. 6. Station rankings based on $S(i)$, $R(i)$, $N(i)$, and GLS(i), using annual peak discharge time series for the period 1952–1998 in southern Illinois. Gaging stations numbers correspond to those in Table 2.

Table 3
Correlation matrix for the rankings of information transmitted, $S(i)$, information received, $R(i)$, net information, $N(i)$, and the generalized least square method, GLS(i), using annual peak flows for the period 1952–1998 in southern Illinois

Correlation coefficient	$S(i)$	$R(i)$	$N(i)$	GLS(i)
$S(i)$	1.0000	0.8881	0.8392	−0.3846
$R(i)$	0.8881	1.0000	0.5664	−0.5175
$N(i)$	0.8392	0.5664	1.0000	−0.2238
GLS(i)	−0.3846	−0.5175	−0.2238	1.0000

programm GLSNET (Tasker and Stedinger, 1989) and uses total net entropy $N(i)$ as a penalty function in place of the cost function. For this analysis all the station costs are assumed equal. Penalty values were calculated using the formula:

$$P(i) = 1 - M \times \mathrm{N}(i) \tag{17}$$

where $P(i)$ is the penalty for station i, M is a variable multiplier to be determined, and $N(i)$ is the total net information at station i. The penalty function was

Table 4
Information transmitted, $S(i)$ information received, $R(i)$ and net information, $N(i)$ and station rankings based on $S(i)$ $R(i)$ $N(i)$ and GLS(i) using annual peak flows for the period 1979–1998 in southern Illinois. Smaller ranks indicate that stations have less regional value

Station no. (i)	USGS station no.	Information			Rank (r)			
		$S(i)$	$R(i)$	$N(i)$	$S(i)$	$R(i)$	$N(i)$	GLS(i)
1	03345500	0.7408	0.6622	0.0786	7	6	11	2
2	03346000	0.7700	0.6397	0.1302	8	4	12	6
3	03378000	0.7154	0.7384	−0.0230	6	9	5	9
4	03379500	0.7853	0.8213	−0.0360	12	12	3	3
5	03380500	0.6585	0.7368	−0.0783	4	8	2	4
6	03381500	0.7702	0.7690	0.0012	9	10	8	8
7	05576000	0.6235	0.6544	−0.0309	3	5	4	7
8	05577500	0.4802	0.5919	−0.1117	2	2	1	10
9	05587000	0.6789	0.6011	0.0778	5	3	10	1
10	05588000	0.7846	0.7184	0.0662	11	7	9	11
11	05594000	0.7793	0.8012	−0.0219	10	11	7	5
12	05597500	0.4730	0.4959	−0.0230	1	1	6	12

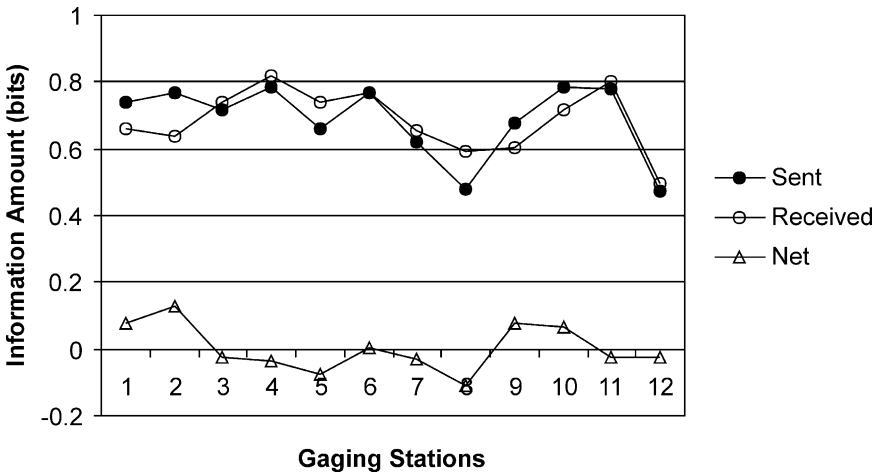

Fig. 7. Information transmitted, $S(i)$, information received, $R(i)$, and net information, $N(i)$, using annual peak discharge time series for the period 1979–1998 in southern Illinois. Gaging stations numbers correspond to those in Table 4.

calculated for various assumed values for multiplier M, and applied to the stations in eastern Illinois, which is presented in Table 7.

With each set of penalties corresponding to a value of M, the program produced station ranks for the combined model. The combined model then was compared with $N(i)$ and GLS through the correlation coefficient between the ranks based on each method. For small M, the differences in penalty for all stations were small, and the resulting ranks of the combined method correlated strongly with those of GLS and had no significant correlation with the ranks based on entropy. Conversely, for large values of M, the ranks of the combined model correlated strongly with those of $N(i)$ and weakly with the ranks of GLS. The 'optimum' M was chosen such that the ranks based on the combined model preserve high correlation with those of the GLS and entropy models (Fig. 9).

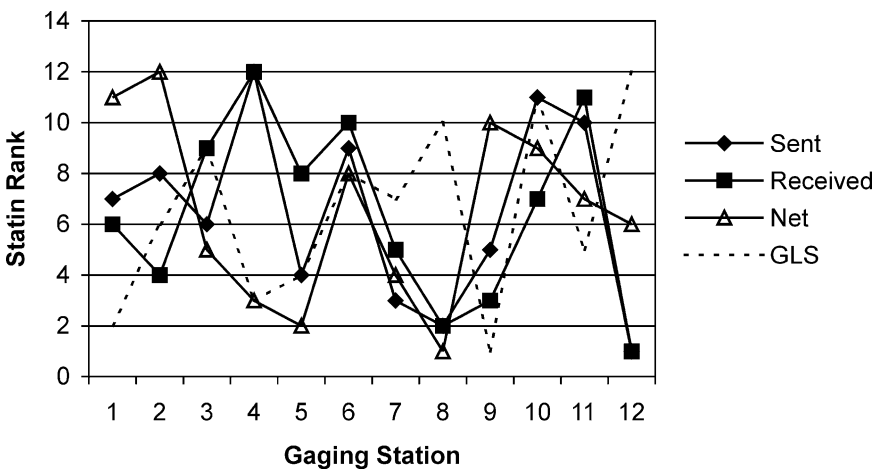

Fig. 8. Station rankings based on $S(i)$, $R(i)$, $N(i)$, and GLS(i), using annual peak discharge time series for the period 1979–1998 in southern Illinois. Gaging stations numbers correspond to those in Table 4.

Table 5
Correlation matrix for the rankings of information transmitted, information received, net information, and the generalized least square method using annual peak flows for the period 1979–1998 in southern Illinois

Correlation coefficient	$S(i)$	$R(i)$	$N(i)$	GLS(i)
$S(i)$	1.0000	0.7622	0.3566	−0.2657
$R(i)$	0.7622	1.0000	−0.1259	−0.2797
$N(i)$	0.3566	−0.1259	1.0000	−0.2308
GLS(i)	−0.2657	−0.2797	−0.2308	1.0000

The range of M between $M = 10$ and $M = 20$ for the combined model maximizes the average correlation. For $M = 10$, the new model correlates equally with GLS and entropy, with the average correlation coefficient equal to 0.73. The average correlation coefficient reaches its maximum at $M = 15$ ($r = 0.76$). The results are the station ranks for all three methods and $M = 15$ are presented in Fig. 10. Some stations have very different ranks when using different methods. For example, station 05580000 has a very high regional value using GLSNET (rank $r = 13$) and a very low regional value using entropy $N(i)$ ($r = 2$). The rank for the same station based on the combined model is equal to 8.

6. Conclusions

This paper presents a newly developed entropy-based method for analyzing the regional value of stations in a gaging network. The method represents an extension of previously published ideas (Yang and Burn, 1994; Chapman, 1986) and was compared with the USGS GLS network analysis. Both methods were tested on a network of gaging stations in southern Illinois, for two time periods, 1952–1998 and 1979–1998, and eastern Illinois for the period of 1952–1998. In addition to the difference between the methods, the station ranks within each method varied significantly with changes in the number of stations in the network.

Entropy-based measures for uncertainty and information transfer are sensitive to the number of class intervals for data. However, the ranks based on entropy do not change much with the number of discrete intervals. In the presented example of using the net information transfer as a criterion for evaluating the stations, the first station to be discontinued retained the same rank for the entire range of class intervals ($K = 5–25$).

While the entropy model was based on the principle of retaining those stations receiving low

Table 6
Station rankings for various methods and two time periods: 1979–1998 (I), and 1952–1998 (II) in southern Illinois

USGS station no.	Information						GLS	
	Transmitted		Received		Net			
	I	II	I	II	I	II	I	II
03345500	9	7	8	6	9	11	2	6
03346000	11	8	11	4	12	12	6	4
03378000	8	6	6	9	10	5	9	5
03379500	12	12	12	12	11	3	3	9
03380500	5	4	10	8	1	2	4	3
03381500	10	9	9	10	8	8	8	7
05576000	3	3	3	5	5	4	7	8
05577500	1	2	1	2	2	1	10	10
05587000	4	5	4	3	6	10	1	1
05588000	6	11	5	7	7	9	11	12
05594000	7	10	7	11	4	7	5	2
05597500	2	1	2	1	3	6	12	11
Correlation between I and II	0.842		0.671		0.532		0.699	

Table 7
Station numbers, entropy, station cost, $C(i)$, as a function of multiplier M and net-entropy $N(i)$, for stations in eastern Illinois for the period 1952–1998

Station	Entropy $N(i)$	Variable multiplier M						
		1	2	3	5	10	15	20
03339000	−0.0318	0.968	0.936	0.905	0.841	0.682	0.523	0.365
05439500	−0.0470	0.953	0.906	0.859	0.765	0.530	0.295	0.061
05520500	−0.0489	0.951	0.902	0.853	0.756	0.511	0.267	0.023
05525000	−0.0845	0.916	0.831	0.747	0.578	0.155	−0.267	−0.690
05525500	0.0103	1.010	1.021	1.031	1.051	1.103	1.154	1.205
05526000	0.0041	1.004	1.008	1.012	1.021	1.041	1.062	1.082
05527500	0.0213	1.021	1.043	1.064	1.107	1.213	1.320	1.426
05552500	−0.0052	0.995	0.990	0.984	0.974	0.948	0.922	0.896
05554500	0.0192	1.019	1.038	1.058	1.096	1.192	1.288	1.384
05556500	0.0247	1.025	1.049	1.074	1.123	1.247	1.370	1.493
05567500	−0.0564	0.944	0.887	0.831	0.718	0.437	0.155	−0.127
05572000	0.0071	1.007	1.014	1.021	1.036	1.071	1.107	1.142
05580000	−0.0751	0.925	0.850	0.775	0.624	0.249	−0.127	−0.502

amounts and transmitting high amounts of information, the GLS-based network analysis uses physical characteristics of the watersheds, geographic location and other feature to produce various statistics. The stations are evaluated and ranked based on those statistics. As a result of different assumptions and different methodologies, the entropy and GLS methods produced different station ranks. The correlation coefficient between the ranks of the two methods suggests that GLS ranks are inversely proportional to the information transmitted, $S(i)$, information received, $R(i)$, and to the net information, $N(i)$. Stations located in an area of high gage density tend to receive and transmit more information. Gages having less significant regional value transmit substantially less information then they receive.

Finally, a hybrid combination of the entropy and GLS measures of regional value of stations was

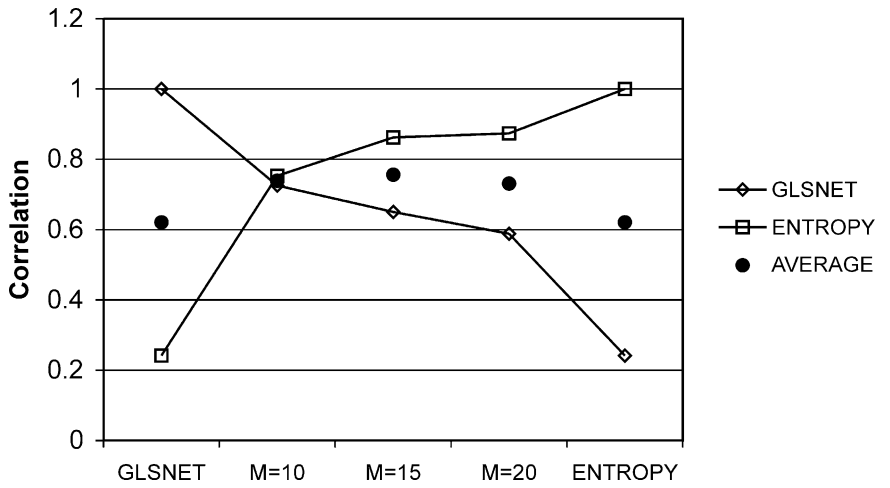

Fig. 9. Correlation between station rankings obtained using the combined model and station rankings based on entropy ($M = \infty$) and GLSNET ($M = 0$) for stations in eastern Illinois for the period 1952–1998.

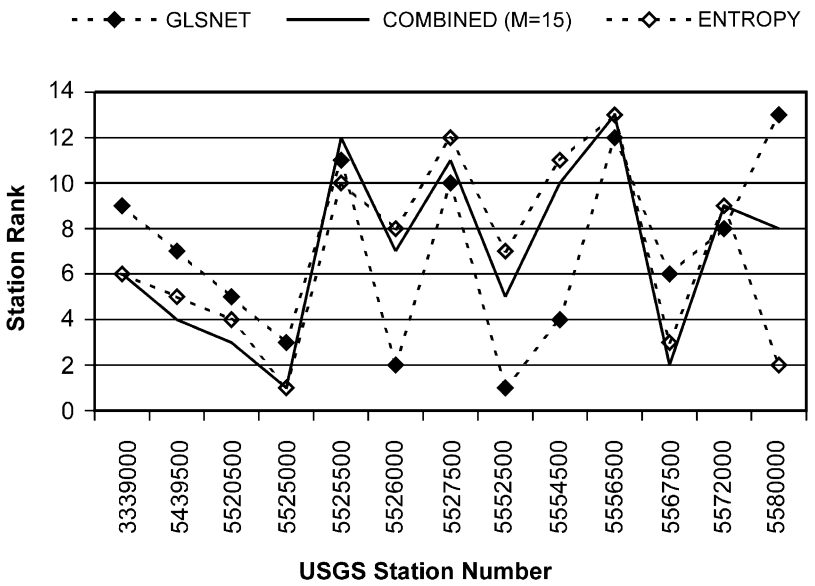

Fig. 10. Station rankings as a function of three chosen methods: entropy, GLS, and combined ($M = 15$) for stations in eastern Illinois for the period 1952–1998.

presented as an attempt to overcome the differences between the two approaches. By a suitable choice of parameter M, the station rankings based on the combined method can preserve correlation with rankings using both entropy and GLS.

Numerous other uncertainties of ranks for all the methods presented here remain to be assessed and quantified. The future research should also focus on refining the methods for estimating parameters K and M. Application of the presented approach to different regions would certainly provide more insight into the ongoing problem.

Acknowledgements

This material is based upon work supported by the Illinois Department of Natural Resources, Office of Water Resources (OWR), under Award No. 09911S99-273. Arlan Juhl, OWR, served as project manager. We would also like to acknowledge the valuable advice of Wilbert O. Thomas, Michael Baker Corporation, Alexandria, Virginia.

References

Amorocho, J., Espildora, B., 1973. Entropy in the assessment of the uncertainty of hydrologic systems and models. Water Resources Research 9 (6), 1511–1522.

Chapman, T.G., 1986. Entropy as a measure of hydrologic data uncertainty and model performance. Journal of Hydrology 85, 111–126.

Gandin, L.S., 1965. Objective analysis of meteorological fields, Israel Program for Scientific Translations, p. 242.

Harmancioglu, N.B., Alpaslan, N., 1992. Water quality monitoring network design: a problem of multi-objective decision making. AWRA Water Resources Bulletin 28 (1), 179–192.

Harmancioglu, N.B., Yevjevich, V., Obeysekera, J.T.B., 1986. Measures of information transfer between variables. In: Shen, H.W., (Ed.), Proceedings of Fourth International Hydrology Symposium—Multivariate Analysis of Hydrologic Processes, pp. 481–499.

Husain, T., 1989. Hydrologic uncertainty measure and network design. Water Resources Bulletin 25 (3), 527–534.

Karasev, I.F., 1968. Principles for distribution and prospects for development in a hydrologic network. Soviet Hydrology 6, 560–588.

Leighton, M.M., Ekblaw, G.E., Horberg, L., 1948. Physiographic Divisions of Illinois, Illinois, US Geol. Surv. Report of Invest. 129, Champaign, II.

Linfoot, E.H., 1957. An informational measure of correlation. Information and Control 1, 85–89.

Mades, D.M., Oberg, K.A., 1986. Evaluation of the USGS Gaging-station network in Illinois, US Geol. Surv., Water Resour. Invest, 86-4088, 88 pp.

Moss, M.E., Gilroy, E.J., 1980. Cost-effective Stream-gaging strategies for the Lower Colorado River Basin: US Geol. Surv., Water-Supply Paper 2178, 111 pp.

Moss, M.E., Karlinger, M.R., 1974. Surface water network design by regression analysis simulation. Water Resources Research 10 (3), 427–433.

Moss, M.E., Tasker, G.D., 1991. An intercomparison of hydrological network-design technologies. Hydrological Sciences Journal 36 (3), 209–221.

Moss, M.E., Gilroy, E.J., Tasker, G.D., Karlinger, M.R., 1982. Design of Surface Water Data Networks for Regional Information, US Geol. Surv. Water-Supply Paper 2178, 33 pp.

National Research Council, 1992. Regional Hydrology and the USGS Stream Gaging Network, National Research Council, National Academy Press, Washington, DC.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1995. Numerical recipes in Fortran 77, the art of scientific computing. Statistical Description of Data, second ed., Cambridge University Press, New York, Chapter 14; pp. 626–630.

Stedinger, J.R., Tasker, G.D., 1985. Regional hydrologic analysis 1: ordinary, weighted, and generalized least squares compared. Water Resource Research 21 (9), 1421–1432.

Stedinger, J.R., Tasker, G.D., 1986. Regional hydrologic analysis 2: model-error estimators, estimation of sigma and log-pearson type-3 distribution. Water Resource Research 22 (10), 1487–1489.

Tasker, G.D., 1986. Generating efficient gauging plans for regional information, Integrated Design of Hydrological Networks, 158. IAHS Publication, pp. 269–281.

Tasker, G.D., Stedinger, J.R., 1989. An operational GLS model for hydrologic regression. Journal of Hydrology 111, 361–375.

Thomas, W.O., Jr., 1994. World Meteorological Organization, Operational Hydrology Report No. 41, An Overview of Selected Techniques for Analysing Surface-Water Data Networks.

USGS, 1999. Streamflow Information for the Next Century, A Plan for the National Streamflow Information Program of the US Geological Survey, USGS Open File Report, 99-456.

Valdes, J.B., Rodriguez-Iturbe, I., Vicens, G.J., 1975. A Bayesian Approach to Multivariate Hydrologic Synthesis, Ralph M. Parsons Laboratory for Water Resources and Hydrodynamics, Report No. 201, School of Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Wahl, K.L., Thomas, W.O., Jr., Hirsch, R.M., 1995. Stream Gaging Program of the USGS, USGS Circular, 1123.

Yang, Y., Burn, D.H., 1994. An entropy approach to data collection network design. Journal of Hydrology 157, 307–324.